4

LEVEL II

12

AD A091681

DTIC ELECTE NOV 14 1980

C

DECISION PROCESSES LABORATORY
UNIVERSITY OF OKLAHOMA

OFFICE OF NAVAL RESEARCH
CONTRACT NUMBER N00014-77-C-0615
WORK UNIT NUMBER NR19[illegible]
REPRODUCTION IN WHOLE OR IN PART IS
PERMITTED FOR ANY PURPOSE OF THE
UNITED STATES GOVERNMENT
APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

80 11 07 026

12

HYPOTHESIS GENERATION: A FINAL REPORT OF THREE YEARS OF RESEARCH

CHARLES F. GETTYS, CAROL MANNING,
TOM MEHLE, AND STAN FISHER

TR 15-10-80 OCTOBER 1980

PREPARED FOR

OFFICE OF NAVAL RESEARCH
ENGINEERING PSYCHOLOGY PROGRAMS
CONTRACT NUMBER N00014-77-C-0615
WORK UNIT NUMBER NR 197-040

DECISION PROCESSES LABORATORY
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF OKLAHOMA
NORMAN, OK. 73019

(9) Final rept. Aug 77 - Aug 80,

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER (14) TR-15-10-80	2. GOVT ACCESSION NO. AD-A091681	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) Hypothesis Generation: A Final Report of Three Years of Research		5. TYPE OF REPORT & PERIOD COVERED Final : Aug. 1977-Aug. 1980
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Charles F. Gettys, Carol Manning, Tom Mehle, ~~and~~ Stanley D. Fisher		8. CONTRACT OR GRANT NUMBER(s) (15) N00014-77-C-0615
9. PERFORMING ORGANIZATION NAME AND ADDRESS Decision Processes Laboratory, Department of Psychology, University of Oklahoma, Norman, Oklahoma 73019		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 197-040 (12) 59
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research, Engineering Psychology Programs-Code 455, 800 N. Quincy Street, Arlington, Virginia 22217		12. REPORT DATE (11) 3 Mar[illegible]980
		13. NUMBER OF PAGES 53
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

1) hypothesis generation 2) decision theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This project was devoted to a study of hypothesis generation. Hypothesis generation is the process by which the decision maker specifies possible hypotheses, or states of the world, that are relevant in a decision problem.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601 |

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

410718

JOB

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

This final report summarizes 14 experiments conducted over a three-year period. First discussed is a hypothesis generation model and research which addresses the model. Several major findings were obtained: 1) Hypothesis retrieval from memory is impoverished. Hypothesis generators are not able to retrieve all relevant hypotheses from memory that should be considered in a decision problem. 2) Hypotheses that are retrieved from memory are first checked for logical consistency with the data. Those hypotheses that are logically consistent may be assessed further for plausibility. 3) Hypothesis generators think that collections of hypotheses which they generated are much more complete than they actually are.

The next section discusses research on hypothesis generation performance. Topics include protocol analysis, group hypothesis generation, the biasing effects of schemata, individual differences in hypothesis generation, and generalizing to expect populations.

A third section is devoted to a survey of research relevant to aiding the hypothesis generation process. An artificial aid for retrieving hypotheses from memory is discussed. Also discussed are other ways of improving hypothesis generation performance.

The general conclusion of this project is that both the failure to retrieve enough hypotheses from memory and the subjects' belief that these collections of hypotheses are more complete than they actually are can be traced to deficiencies in the memory retrieval process.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

Table of Contents

Acknowledgments

This is the final report of a three-year program of research sponsored by the Office of Naval Research, Engineering Psychology Programs. The contract between the Office of Naval Research and the University of Oklahoma was N00014-77-C- 0615, work unit number NR197-040, Charles F. Gettys, principal investigator.

We wish to thank Martin A. Tolcott, director of the Engineering Psychology Programs. His role as a facilitator of our research was invaluable, and he created the ideal conditions under which we could explore our ideas and learn from our mistakes.

Various individuals contributed to this program of research. Most of the research to be reviewed here was conducted by Stanley D. Fisher, Thomas Mehle, Carol Manning, Suzanne Baca, and Nancy Shelton. As graduate students at the University of Oklahoma, they spent countless hours conducting and implementing this research. Any successes that have been achieved are largely due to their inspired and painstaking efforts.

Undergraduate researchers who worked on this project were Jeff Casey, Dee Ann Fisher (nee' Patterson), Jim Laughlin, Frank Savala and Steve Holmstrom. Their efforts, as unpaid volunteers, are especially appreciated.

We would like to thank Dale Umbach and Bradford Crane for their help in mathematical proofs, and Chuck Rice and Jim White of the University Computer Services for software and consulting services, respectively.

Tom Harrison and the staff of the Office of Grants and Contracts provided editorial and secretarial assistance for which we are most grateful. Glenda Smith and the Psychology Department office staff made substanial secretarial contributions, and Dick Rubrecht cheerfully repaired our computer "one more time."

We wish to thank Alan Nicewander who collaborated with us on one of our studies, and critically read several of our manuscripts. We also wish to acknowledge the help of Rick Acosta, Steve Smith, and Vesta Gettys who read and commented on several of our manuscripts, and freely contributed their ideas.

Our interest in this topic can be traced to conversations between the senior author and Mel Moy of the Navy Personnel Research and Development Center on the topics of threat detection and crisis prevention. These discussions lead to the conclusion that it would be profitable to model the hypothesis generation process.

In recognition of their many contributions to this program of research, Stan Fisher and Tom Mehle are listed as authors. However, they did not have an opportunity to contribute to the actual writing and editing of this report, and are not responsible for any inaccuracies or errors that may be present.

Introduction

This is the final report for the project "Data Plausibility and Hypothesis Generation" sponsored by the Engineering Psychology Programs, Office of Naval Research. The project began August 15, 1978 and ended August 14, 1980. The goal of this project was to develop a model of the hypothesis generation process, and to do research to investigate this model and the hypothesis generation process in general. The strategy employed in this project was to blend concepts drawn from three areas: decision analysis, behavioral decision theory, and cognitive psychology. As part of this project, 15 experiments were conducted, and 9 technical reports were issued concerning the process of hypothesis generation.

This report is organized as follows: The final form of the hypothesis generation model which evolved from this program of research is discussed first. This section deals with the research relevant to the hypothesis generation model. In a second section, research addressing other more general aspects of hypothesis generation is discussed. A third section discusses applied research which investigated possible ways of improving hypothesis generation. Finally, an overview which gives the most important conclusion that can be drawn from this research is presented.

The discussion that follows is organized according to topics and does not attempt to explain experimental procedures and results in detail. To attempt this task would result in several hundred pages of text that would be redundant with our previous technical reports. Instead, as various topics are discussed, reference is made to previous technical reports which contain these details, or to reports which contain relevant references to the general literature. So that interested readers can obtain more information, these technical reports are cited using numerals (ie. 1, 5, 9), and particularly relevant reports which contain our most recent or complete treatment of a given topic are underlined (ie. 2, <u>5</u>, 7).

A Hypothesis Generation Model and Related Research

The hypothesis generation task

Problem structuring is a predecision process by which the decision maker develops the salient characteristics of the decision problem. The decision maker must first develop the objectives and constraints of the decision problem. Once the over-all objectives are formulated, various structural elements are supplied. Structural elements may include: possible acts which are specified by the decision maker, relevant states of the world (hypotheses), and possible outcomes. Outcomes are determined by the both the act that the decision maker chooses and the state of the world that obtains when that act occurs.

This project was devoted to the study of hypothesis generation, i.e., the process by which the decision maker generates the relevant states of the world. In terms of problem structuring, the decision maker should be able to generate the possible states of the world that may affect the outcomes of any acts that are taken. For some problems this task may be easy. The decision maker may generate hypothesized states of the world related to a problem which has been experienced before. In these situations possible hypotheses may be readily retrieved from memory because they are few in number and routine in nature. Another important class of problems exists where hypothesis generation is a crucial component of problem structuring. Examples of tasks which require hypothesis generation include medical diagnosis, automotive and electronic trouble shooting, and the scientific process itself. Tasks in this category are particularly difficult to solve when the number of possible hypotheses is large and the decision maker cannot rely on past experience to narrow the field to several obvious hypotheses. It is particularly important that the decision maker include the actual state of the world in the problem structure, because any subsequent decision that fails to consider that state of the world may be wrong. For example, if your auto mechanic fails to entertain the hypothesis that a dirty carburetor is responsible for your car's bad performance, you may pay for a series of adjustments or part replacements that do nothing to correct the problem. Similarly, if your doctor fails to consider the disease that you actually have, the whole treatment regime may be inappropriate, or even dangerous to your health. Therefore, one important part of the hypothesis generation task is the inclusion of the true state of the world in the set of possible hypotheses. It is important that the set of hypotheses generated by the decision maker should be as complete as possible. Ideally, the set should be exhaustive; however, a practical decision maker usually neglects improbable hypotheses because these states of the world appear so unlikely that they can safely be neglected.

The hypothesis set that the decision maker creates should contain plausible hypotheses. The construct of "plausibility" includes the notion that for a hypothesis to be included in the set of hypotheses it should be sufficiently probable to be worth further analysis. This does not necessarily involve an assessment process as detailed and thorough as is typically implied by the term "probability assessment." All that is logically necessary at the early stages of problem structuring is that the decision maker make a rough "go/no go" decision in regard to each hypothesis. Hypotheses that pass this crude plausibility test may be more carefully assessed in later stages of decision analysis. While it is possible that plausibility assessment and probability assessment share common elements, there are a few clear differences. The first major difference is in the nature of the task requirements. In a probability assessment task, assessments are usually made about the relative likelihood of a set of specified hypotheses known to the decision maker. In a hypothesis generation task, hypotheses are evaluated with respect to whether or not they should be considered further. This evaluation is complicated by the fact that the evaluation should be relative to both previously-specified hypotheses that the decision maker may have and unspecified hypotheses that are yet to be generated by the decision maker. These task differences suggest that calling the process of deciding if a hypothesis should be included in the set of hypotheses "probability assessment" may be misleading because of the task differences between the two processes. We do not know at this time if the same psychological processes are used in both types of assessment, although it seems quite certain that both processes share common elements.

Hypothesis generation tasks also have the characteristic that generated hypotheses should be consistent with any available information. This information may be specific data or knowledge about the task. Obviously, hypotheses that are inconsistent with the available evidence should not be considered. Information provided by data and the task has a second important role, since it serves as a basis for the memory search processes described in the next section. Although the emphasis will be on memory search processes, the importance of the data as constraints to the logical possibility of hypotheses should be kept in mind.

The hypothesis generation process could operate in a number of different ways depending on the task requirements. For example, during a "brain-storming" session, decision makers may be asked to generate any hypotheses that come to mind irrespective of their plausibility or implausibility. In another situation, the decision maker's task may be to generate all hypotheses that are logically consistent with the data, even though some of the hypotheses are unlikely. In a third situation, the decision maker's task may be to generate a set of plausible hypotheses and to be concerned with whether or not each hypothesis in that set is sufficiently plausible to be included as a candidate for subsequent decision analysis.

Overview of the hypothesis generation model

The hypothesis generation model that has been developed as part of this project has three components or subprocesses. The first subprocess is an executive process. The executive subprocess controls hypotheses generation according to the demands of the task. It initiates memory searches and controls plausibility assessment. The memory search subprocess is responsible for both retrieving hypotheses from memory, and for furnishing information necessary for plausibility assessment. The third subprocess is that of plausibility assessment. In this subprocess hypotheses may be checked to see if they are logically consistent with the data. More sophisticated plausibility judgments may also be made. The plausibility assessment subprocess decides if a hypothesis is sufficiently plausible to warrant further processing. Figure 1 shows this model in summary form. In the three sections that follow, each of the subprocesses and their experimental results are discussed.

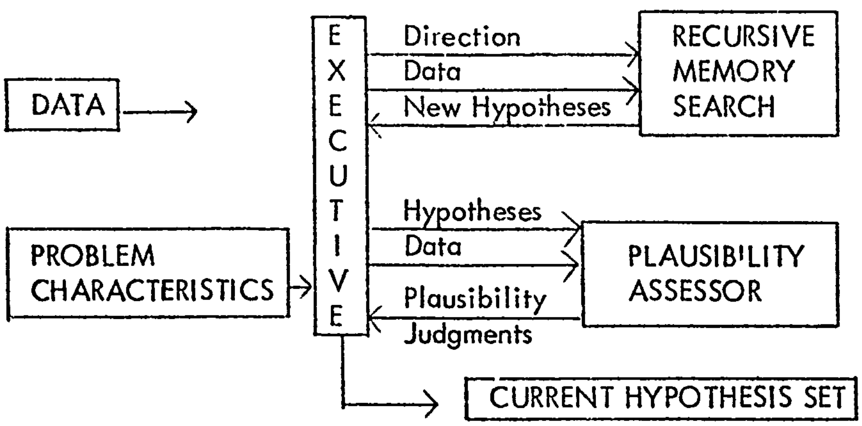

Figure 1. Major subsystems in hypothesis generation model.

Hypothesis retrieval from memory

When the hypothesis generation process begins, the decision maker has an empty hypothesis set which must be populated. A reasonable goal is to develop a set of hypotheses that is as complete as possible. To accomplish this end, hypotheses must be retrieved from memory. The model assumes that available data and other task information are used to search memory. Memory is assumed to be organized in a semantic net (1, 3). Searches are made for each datum. If a hypothesis consistent with the available data is encountered in this search process, then it is tagged in memory to reflect this encounter. When a hypothesis accumulates a critical number of tags, the executive notes this fact, and the hypothesis is retrieved from memory for further processing. A detailed discussion of the memory search subprocess has been provided (1), but some of the results obtained during an evaluation of the model are of greater interest.

The first point of interest is whether or not the search and retrieval process produces candidate hypotheses which are logically consistent with all data. An analysis of the hypothesis generation task suggests that this should be a minimum requirement of any hypothesis included in the final hypothesis set. When does consistency checking occur? Does the memory search subprocess necessarily produce hypotheses that are logically consistent with all data or is consistency checking performed after retrieval from memory? Perhaps a hypothesis must be tagged by all data before it is retrieved by the executive. One assumption of this version of the model is that a hypothesis would not receive a tag from a datum if it is inconsistent with that datum. In a second version of the model it might be assumed that any hypothesis encountered in the memory search may be retrieved for further processing. Under this assumption, retrieval could follow from a single tag.

The "one-tag" version and the "all-tag" version are limiting cases of the tagging model. A task analysis suggested that it was unlikely that the "one-tag" version would be correct. If a hypothesis suggested by any of the data is retrieved for futher processing, then using the "one-tag" version, the decision maker would have to process a large number of hypotheses most of which would be inconsistent with one or more data. If, however, all hypotheses suggested by the data had to be tagged by all data, then the decision maker would retrieve very few hypotheses, and would probably fail to retrieve many relevant hypotheses. It seems reasonable to assume that the decision maker should choose a strategy that lies somewhere between these two extremes.

The tagging model was designed so that the criterion number of tags was a free parameter, and it was used as a measurement tool to address this issue. A study (1) was conducted where decision makers retrieved hypotheses from either a set of six data, or subsets of these data which consisted of three data, or only one datum.

The criterion number of tags for retrieval to occur was estimated from these data, and was found to be between two and three. Recently, we have shown that this conclusion does not depend on the assumptions of the tagging model; other similar models would yield the same conclusions.

The major implication of this result is that hypotheses are retrieved from memory using two or three data as retrieval cues. Therefore, retrieved hypotheses are at least partially consistent with the available data. These results also suggest that the memory search process may produce hypotheses that will be discarded in subsequent assessment because they are not logically consistent with the rest of the data.

A second point of interest deals with the efficiency of the hypothesis retrieval process. In order to study this process, the retrieval performance of the subjects was compared to a "minimally-adequate hypothesis set" developed by the experimenters. This minimally adequate hypothesis set consisted of the three most-plausible hypotheses which the experimenters felt should be included in an "adequate" set of hypotheses generated by the subjects. The set for each problem was chosen conservatively and many other plausible hypotheses were excluded. Only 19.9% of the subjects were able to retrieve these three hypotheses. We also explored the effect of relaxing the definition of adequate performance. We found that 50% of the subjects were able to retrieve two out of three of the "minimally adequate" hypotheses, while 92% of the subjects were able to retrieve one of the three. This result was our first indication that the hypothesis generation process was less than adequate, and it has been replicated many times using more objective criteria of performance. Similar results are discussed in a later section of the paper. The results discussed here are important because they suggest that the memory search process is involved in the deficiencies in hypothesis generation reported throughout this project.

Checking hypotheses for logical consistency

Results from the tagging study (1) of the memory search model suggest that the decision maker will often retrieve a hypothesis from memory using several data. This newly-retrieved hypothesis may or may not be consistent with all of the remaining data that were not used in its retrieval. A consistency checking process may exist in which the decision maker checks the newly-retrieved hypothesis for logical consistency with any remaining data. Such a process should be relatively fast, as compared to hypothesis retrieval. Using the hypothesis as a retrieval cue, the decision maker should perform a high-speed memory scan to examine whether the hypothesis is consistent with the remaining data. For reasons of efficiency, the consistency checking process should be self-terminating, ie. the consistency checking should stop if a datum is encountered which is inconsisent with the newly-retrieved hypothesis. If a hypothesis passes this consistency check, then it is logically consistent with all of the data, and it has met the minimum plausibility requirements. Plausibility assessment may stop at this point, or it may continue, depending upon the demands of the task.

A series of experiments (3) was conducted to investigate the nature of consistency checking. The first experiment asked whether or not consistency checking exists. Subsequent experiments were conducted to examine the speed of consistency checking relative to hypothesis retrieval, and whether or not consistency checking is a self-terminating process.

The first experiment was an attempt to demonstrate that consistency checking exists. An instructional manipulation was used in which subjects were instructed to either respond with the first hypothesis that occured to them, irrespective of its consistency, or were instructed only to respond with a consistent hypothesis. Hypothesis generation problems containing various numbers of data were used. We predicted an interaction between the time necessary to generate a hypothesis in the two conditions and the number of data in the problem. While large differences were observed between the two conditions, the interaction was not significant. We believe that the inconclusive results of this experiment were due to the subjects' inability or unwillingness to respond with the first hypothesis that occurred to them even though they were instructed to do so.

In a study which was too recent to be discussed in the original technical report (3), the question of the existence of consistency checking was investigated again. In this study a somewhat different approach was used. Subjects were asked to generate consistent hypotheses in response to data. Immediately after they generated a hypothesis, they were shown a list of inconsistent hypotheses that had been generated by another group of subjects.

Subjects scanned the list of inconsistent hypotheses, and identified any that had "crossed their minds" during hypothesis generation.

It was estimated that subjects retrieved an average of 1.83 inconsistent hypotheses before they retrieved their first consistent hypothesis. This experiment contained a manipulation to control for the obvious demand characteristics. Subjects may have picked hypotheses from the list to please the experimenters. It is unlikely that these results could be explained in that way. It was concluded that subjects do check newly-retrieved hypotheses for consistency, and that inconsistent hypotheses are discarded at this time. These results also add support to the conclusion that memory is searched using only part of the available data. The memory search result implies that inconsistent hypotheses are retrieved from memory, and this consistency checking experiment demonstrated that inconsistent hypotheses are retrieved from memory and are then discarded.

The next experiment in this series (3) addressed our prediction that consistency checking is a more rapid process than hypothesis retrieval. Two experimental conditions were compared. Subjects in condition one generated hypotheses in response to varying amounts of data. Subjects in condition two were given the hypotheses that the first group had generated, and were asked to check them for consistency using the same data. Using a Sternberg memory search procedure (3), the time to process each additional datum was estimated. Subjects who generated hypotheses took 1.8 seconds per datum, while consistency checking subjects were able to process each datum in .7 second, i.e. between two and three times faster than hypothesis generation subjects.

The final experiment in this series examined the self-termination prediction. Subjects were provided with a hypothesis and were asked to check three-data problems for consistency with respect to that hypothesis. The position of a disconfirming datum in the data set was varied for problems where the hypothesis was inconsistent with the data. Subjects responded faster when the disconfirming datum was earlier in the sequence of data than when it was later. This result is consistent with a self-terminating process.

The results of the experiments investigating the existence of consistency checking suggest that subjects retrieve hypotheses which are found to be inconsistent with a set of data. We believe that consistency checking occurs in the hypothesis generation process and that subjects tend to retrieve hypotheses in response to only part of the available data. Thus, the results support the predictions of the partial-retrieval consistency checking model of hypothesis generation rather than the alternate retrieval model which assumes that subjects retrieve consistent hypotheses using all data as retrieval cues.

The results of experiment two of this series demonstrated that less time is needed to process an additional datum during consistency checking than during hypothesis retrieval. These results are consistent with the predictions based upon the search properties of hypothesis retrieval versus the verification properties of consistency checking. Experiment three of this series provided evidence that consistency checking is a self-terminating process.

These results are important for an understanding of the hypothesis generation process. They more clearly define the role of memory in hypothesis generation, and the processing of hypotheses subsequent to retrieval from memory. These results, when combined with our other research, are consistent with the following model of hypothesis generation:

Hypotheses are retrieved from memory using several data. If the data are numerous, then retrieval is based upon only a part of the available data. Upon retrieval, hypotheses are checked for logical consistency with any remaining data using a high-speed semantic verification process. If a logical inconsistency is found between a hypothesis and a datum then processing stops, and the hypothesis is labeled as inconsistent. If, however, the hypothesis survives the consistency checking process, then further processing can occur depending on the task demands. The consistency checking process is faster than the retrieval process because retrieval involves a search for hypotheses that are suggested by several data; whereas, consistency checking involves verifying semantic relationships among a hypothesis and data that are already active in memory.

Hypotheses that survive the consistency checking process have met the _minimal_ task requirement for hypothesis generation, that of logical consistency with the data. They are not necessarily plausible hypotheses; plausibility can be established by further processing if the task requires this type of assessment.

Our use of the term "consistency checking" has been solely confined to high-speed semantic verification. We do not intend to imply that other processes which might be called "consistency checking" do not exist. Thus, a scientist may spend months determining if a hypothesis is consistent with data. This is not the process studied here, and this distinction becomes clearer if a scientist's work is termed "hypothesis assessment." We have studied the early phases of the hypothesis generation process, and we believe that in the first few seconds of hypothesis generation a hypothesis is retrieved from memory using part of the data and then checked for consistency with the remainder of the data.

Plausibility assessment of generated hypotheses

After a hypothesis is retrieved from memory and checked for logical consistency, further processing may occur to determine if the hypothesis is sufficiently plausible to be included in the set of hypotheses that the decision maker is entertaining. Secondly, the decision maker must decide if more hypotheses should be included in the set of hypotheses, or if the set is complete enough to be satisfactory. Once the set is sufficiently populated with hypotheses, attention can be turned to other aspects of problem structuring. This task analysis suggests that the decision maker should have some sensitivity to the plausibility of both individual hypotheses and the collection of hypotheses called the hypothesis set.

As discussed previously, the task of estimating the plausibility of hypotheses is somewhat different than a probability or odds estimation task. The task of the decision maker in hypothesis generation is to populate an empty hypothesis set; whereas, in probability or odds estimation the task is to estimate the relative likelihood of an existing set of specified hypotheses. The probability estimator, for example, need only be concerned with the relative likelihoods of a set of enumerated hypotheses. The hypothesis generator, on the other hand, must judge a specified hypothesis that has just been retrieved from memory against a diffuse unspecified set of hypotheses that potentially might be included in the hypothesis set. Before the plausibility of a hypothesis can be established, it must be compared to other alternative hypotheses which may or may not be available in memory. Thus, plausibility assessment would seem to be much more formidable than probability or odds estimation, and one might naturally expect that subjects' plausibility assessments will be found less accurate. This kind of judgment is analogous to the difference between absolute and relative judgments in perception where it is commonly known that relative judgments are easier to make than absolute judgments. The plausibility assessor may be making a judgment about a hypothesis in the absence of other hypotheses. As the hypothesis set becomes more populated, plausibility and probability assessment become more similar in nature, and for fully-populated sets the tasks become identical. The same argument holds for judgments of collections of hypotheses where the task is to generate a set of hypotheses which is as complete as possible. Decision makers should continue to generate hypotheses until they believe that the collection of specified hypotheses equals the set of all possible hypotheses.

The first research concerned with hypothesis assessment was an early study done by Gettys and Fisher (cited in 7) which was not a formal part of this project. This study was devoted to the executive control of the hypothesis generation process, and it investigated the rules for deciding if a particular hypothesis or hypothesis set is plausible. Of particular interest in this study was the relationship between these rules and the memory search process. It was found that additional hypotheses were most often generated when data were presented which disconfirmed the set of currently-held hypotheses. The data were examined to see if a fixed criterion of plausibility was used to admit a newly-generated hypothesis to the current set of hypotheses. No evidence for such a fixed plausibility threshold was found. Instead, subjects seemed to be admitting hypotheses into the set only if they were close competitors with the most plausibile hypotheses that had already been generated. This behavior was characterized as a search for "leading contenders" rather than a search for an exhaustive set of hypotheses.

The first study in this project examined the question of whether or not subjects could evaluate the plausibility of hypotheses. Of interest were the plausibility estimates subjects made concerning sets of hypotheses differing with respect to plausibility or completeness. Subjects were given sets of hypotheses which varied in plausibility, and were asked to judge both the plausibility of each hypothesis individually and the collection of hypotheses. The judgments included estimates of the plausibilities of both specified hypotheses and the diffuse set of unspecified hypotheses. These judgments were evaluated by comparing them to a probabilistic model developed for this purpose.

The task which was modeled was that of generating possible academic majors for a hypothetical student at the University of Oklahoma. The hypotheses to be generated were based on the courses the student had taken. The enrollment records for all students currently enrolled in the University were used to determine the probabilistic relationships beween majors and courses. A total of 166,858 enrollment records were tabulated to obtain the posterior probabilities of various majors given selected courses. These veridical values were compared to subjects' estimates to address the accuracy of calibration. This task had the necessary characteristic that the veridical relationships between majors and courses were known, and the task also had the property that most student subjects understood it intuitively. However, it should be noted that many of the relationships between courses and majors are complicated. Students enroll in a program of study for many complex reasons, including personal preference, advice from other students and advisors, and College and University requirements.

In the first experiment (1), subjects estimated the plausibility of three specified hypotheses and a diffuse catch-all hypothesis of "all other hypotheses". They also estimated the plausibility of the specified collection of hypotheses versus the catch-all set. Two major results were obtained. First, as might be expected from the task analysis, plausibility estimates were quite variable, and were only weakly related to the veridical probabilities. Second, the overwhelming majority of these estimates were excessive in respect to the veridical probabilities. Both results were quite reliable, and have since been replicated in several situations (2,7).

It occurred to us that the explanation for this excessive certainty might be that the decision maker must populate the complementary set of unspecified hypotheses before the specified hypotheses (or sets of specified hypotheses) can be assessed accurately. We also had reason to believe that the retrieval of hypotheses from memory was impoverished. If this were the case, then attempts by the decision maker to populate the unspecified set of hypotheses would be only partially successful. Consequently, when plausibility estimates were made, the unspecified set of hypotheses was incomplete; hence, its plausibility was under-estimated. If the plausibility of the unspecified set was under-estimated, then the plausibility of the specified set was necessarily over-estimated.

The next study (2) was a test of this explanation. There were three groups of subjects in this study. One group was essentially a replication of one of the conditions of the previous study. Subjects estimated the plausibility of sets of specified hypotheses and the unspecified catch-all hypotheses much as before. In the other two groups, however, manipulations were introduced which were designed to increase the availability of hypotheses in the catch-all set. In one condition, subjects were encouraged to explicitly populate the catch-all set. This manipulation was chosen because it was believed that asking the subjects to make a formal search of memory for hypotheses would increase the number of "unspecified hypotheses" available in memory. The second manipulation consisted of showing the subjects exemplar hypotheses from an experimenter- generated catch-all set. This manipulation should also increase the availability of hypotheses in the catch-all set.

Both conditions which were designed to increase the availability of hypotheses in the catch-all set produced estimates that were less excessive. Therefore, we concluded that at least part of the excessiveness in plausibility assessment was due to the limited availability of hypotheses in the catch-all set.

Our studies up to this time had used only sets of hypotheses supplied by the experimenter. We were forced to used experimenter-supplied sets because of limitations in the software which determined the probabilistic relationships between courses and majors. We developed an algorithm which would efficiently process the 166,858 enrollment records for all courses and all majors. Then we were able to run a new study which both replicated the previous studies using experimenter-supplied hypotheses, and also allowed us to study plausibility estimates for subject-generated hypotheses. Therefore, one comparision in this study was between experimenter-supplied and subject-generated hypotheses.

Previous studies employed a response mode which was a variant of the odds estimation technique. A direct probability estimation response mode was compared to the odds response mode. The motivation for this manipulation was to make sure that the excessiveness in plausibility estimates was not due to the response mode.

The results replicated our previous research and reinforced our conclusions. Plausibility estimates were excessive for both experimenter-supplied and subject-generated hypotheses. We had predicted that this would be the case because subjects should have difficulty populating the unspecified set of hypotheses in either condition. Somewhat to our surprise, however, subjects who generated their own hypotheses were significantly more excessive than subjects who worked with experimenter-supplied hypotheses. One possible explanation for this effect is that subjects who generated their own hypotheses nearly exhausted their set of plausible hypotheses in populating the specified set, and consequently did a poorer job of populating the unspecified set.

In both response mode conditions excessive estimates were found, although the subjects in the direct probability estimation condition were somewhat less excessive than subjects in the odds estimation condition. (This study was not issued as a technical report because it was a follow-up study for the availability study (2), but was included in the journal version of the availability study.)

Perhaps the most robust and important conclusion that can be drawn from the last three studies is that plausibility estimates of hypotheses are excessive, and that this behavior can be traced to deficiencies of the hypothesis retrieval process.

Research on Hypothesis Generation Performance

Some of the research on hypothesis generation was addressed to a variety of topics including protocol analysis, group processes, the importance of schemata in hypothesis generation, individual differences in hypothesis generation, and the role of expertise in hypothesis generation. Summaries of the important results on these topics are presented in the following section.

Protocol analysis of hypothesis generation

Mehle, in a doctoral dissertation (7), took a rather different approach to the hypothesis generation problem. Using a modification of Simon's protocol analysis technique, the hypothesis generation performance of expert and non-expert auto mechanics was studied in an automotive trouble-shooting task. This study used markedly different research strategies than the other studies in this project, and it independently confirmed several of the observations that were made using more traditional techniques.

Subjects in the protocol analysis task were either undergraduates who professed some knowledge of cars, or expert auto mechanics from the University motor pool. Subjects were given a written description of a malfunctioning automobile, and were asked to "think out loud" while generating hypotheses about the cause of the malfunction. Examination of the protocols revealed evidence for consistency checking. Hypotheses were generated, and then subsequently ruled out as inconsistent with the data.

In addition to the protocol analysis, both the number of hypotheses that the subjects generated were analyzed, and the plausibility estimates for collection of hypotheses that the subjects generated were analyzed. Experts and non-experts generated approximately the same number of hypotheses; the mean number of hypotheses generated per problem was 3.43 and 3.36 for the non-experts and experts, respectively. These means can be compared to the number of hypotheses that were logically possible for the problems. Information provided by the subjects was used to make this estimate in the absence of a completely authoritative source for this information. The hypothesis set for each subject was pooled with that of the other subjects by taking the union of all hypothesis sets. Illogical hypotheses were discarded from this pool (an average of .1 hypotheses per subject per problem). The number of hypotheses in the pooled set is actually a lower-bound estimate of the number of logically-possible hypotheses. The obtained pooled sets contained an average of 17.8 hypotheses per problem. By applying a mathematical model to this situation, Mehle was able to estimate the number of hypotheses that were logically possible was 21.5 hypotheses in the average problem. Thus the average subject was generating approximately 19% of the logically-possible hypotheses per problem. It was impossible to determine if the hypotheses generated by the subjects were implausible or plausible, but subjects' hypothesis sets certainly lacked the desirable characterstic of completeness.

The plausibility estimates of the sets of hypotheses generated by the subjects were also examined. There were no veridical probabilities for this task, but it was possible to exploit the fact that the sum of the probabilities of an exhaustive set should be one. The hypothesis generators in this experiment generated incomplete, impoverished sets of hypotheses. If all subjects' probability estimates are assumed to be true and if these estimates are assigned to the hypotheses in the pooled set, then a probability measure of 5.04 must be assigned to the more complete set of hypotheses developed by pooling. This measure would have been 1.00 had the whole group of subjects been veridical estimators. Thus, subjects were clearly excessive in this task. This result generalizes our earlier conclusions considerably, as it shows similar behavior in a task that was quite different from the "majors from classes task."

In summary, the protocol analysis study, while done in the same laboratory, reached much the same conclusions as other research conducted using different techniques. The data suggested that subjects were impoverished hypothesis generators whose plausibility estimates were excessive.

Group hypothesis generation

One strategy that has frequently been used to improve problem solving performance is to work in small groups rather than as individuals. The mounting evidence that individual hypothesis generators produced impoverished hypothesis sets suggested that it might be profitable to investigate group hypothesis generation to determine the improvement that working in a group affords. In this study (9), subjects either generated majors from classes as individuals, or as a member of an interacting group of four subjects. The pooling technique was used again, but in this case the veridical posterior probabilites of majors given classes were available, and were used rather than a count of logically-possible hypotheses. Thus the posterior probability of hypothesis sets generated by either individuals or small groups could be calculated. It was also possible to calculate the posterior probability of pooled hypothesis sets for artificial groups of various sizes by using Monte Carlo techniques. The function that was obtained from these calculations increased monotonically with group size and usually asymptoted between group sizes of fifteen and twenty. This function can be used to estimate the size of the synthetic group that would have the same performance as an interacting group of size four.

The mean probability of the hypothesis set for individuals was .335 while interacting groups of four had a mean probability of .427. The means reported are the probabilities that the hypothesis sets contained the "true" hypothesis. Thus, as one might expect, group performance is superior to individual performance. However, both individuals and small groups were impoverished hypothesis generators. Although subjects in this task were told to neglect very unlikely ($p<.02$) hypotheses, and so could not be expected to have hypothesis sets with a probability of 1.00, there is ground for much improvement in these performances. A synthetic group of 1.8 individuals was calculated to be equal in performance to an interacting group of four individuals. The hypothesis set probability for a synthetic group of four individuals was .540. Evidently the social interaction in the real group impairs performance by producing a lower performance than would be expected from sharing hypotheses mechanically, as is done in a synthetic group.

These results suggested a general way of examining at least two factors which affect group performance. One factor is the potential increase in information that the group provides. The adage, "Two heads are better than one," has validity in this sense. As group size increases, the amount of new information added by each new member should become less, but the total information possessed by the group increases. The pooling process described earlier is one way to measure the information possessed by the group, and it provides a natural metric for expressing how the amount of task-relevant information increases as group-size increases. The second major factor in interacting groups is the social interaction which occurs. Social interaction may be facilitative, but it is usually found to inhibit group performance (9). When the performance of individuals, synthetic groups, and interacting groups are compared, it is possible to partition performance into an informational component and a social component. In the present experiment, the information that could be gained from pooling the information of four individuals is estimated to be a .205 increment in hypothesis set plausibility (.540 -.335 =.205). Social interaction, however, caused a decrement in performance of .113, as calculated from differences in performance of the interacting and synthetic groups (.427 -.540 = -.113). The actual gain in performance of an interacting group over an individual is .092, and this difference results from the additive combination of informational and social factors.

These ideas allow the researcher in group processes to better understand the results of group research. Differences between interacting groups are difficult to understand because groups differ from individuals both in the amount of information possessed and in social interaction. By partitioning performance into two components, the relative contribution of each component to performance can be better understood.

Schemata in hypothesis generation

One informal observation that we made in several studies was that our subjects appeared to be blind to certain classes of hypotheses. When asked to generate hypotheses, subjects sometimes generated hypotheses that seemed to be related to an implicit interpretation of the data. Other subjects seemed to adopt different interpretations of the data, and to generate a different set of hypotheses. This observation suggests that sometimes interpretations of the data influence the memory retrieval process, thus biasing the subjects toward one type of hypothesis and against another type. This general phenomenon has received some attention in cognitive psychology. The organization of data into a meaningful pattern by making inferences about their meaning is termed a schema in cognitive literature.

When the hypothesis generator is attempting to add hypotheses to a set of hypotheses that have already been suggested, schemata might be expected to play an important role. This situation may occur when the hypothesis generator "inherits" a decision problem. As scientists we are constantly faced with inherited hypotheses which may bias our interpretation of the data and our generation of new hypotheses. Often "inherited" hypotheses suggest particular interpretations of the data which might seem forced in the absence of these hypotheses. In our natural desire to obtain closure, we may accept certain interpretations which relate data to hypotheses. These interpretations may come to represent the data and may even be encoded in memory in lieu of the data. When we attempt to generate new hypotheses, the schema that organized the data may be used instead of the data in searching memory. To the extent that this happens, the hypothesis generation process may be biased.

A study was performed to investigate these ideas and to propose a partial cure for any such tendencies on the part of the hypothesis generator. In this study subjects were given several ambiguous data which could be interpreted by using several schemata. The existence of an "inherited" hypothesis was simulated in some conditions by giving the subject one of several hypotheses to evaluate. These hypotheses were good exemplars of several different schemata that could be used to explain the data. The problems involved generating possible hypotheses about an unknown geographical area known as "X". For example, subjects in one problem were told that one hypothesis that was consistent with area "X" was a bakery. Available data were that 1) Most people spend only a short time in area X, 2) Area X contains unusual smells, and 3) Area X is only open during business hours. Subjects who "inherited" the "bakery" hypothesis were more likely to generate hypotheses such as "restaurant," "fruit stand," or "flower shop." Other subjects were given this same problem but "inherited" the hypothesis "dump" rather than "bakery". These subjects were more likely to generate different hypotheses such as "chemical plant," "sewer treatment plant," or "public restroom." The two schemata that these two hypotheses suggest are "pleasant" and "unpleasant" areas, respectively. Subjects adopting the "unpleasant" schema might reason that people spend as little time as possible in dumps because dumps smell bad, and so are unpleasant places. Many dumps are supervised, and hence are only open during business hours. Consequently, subjects might tend to search memory for other similar unpleasant places that have bad smells and are open only during business hours. "Bakery" subjects, on the other hand, may reason that bakeries smell unusual but pleasant, serve their customers quickly, and are open during business hours. These subjects should be biased to search memory for other businesses that have unusual but pleasant smells. In a third condition, subjects were given no inherited hypothesis. All subjects were encouraged to generate as many hypotheses consistent with the data as possible.

As might be expected, these schemata differed in accessibility. Subjects in the "no hypothesis" condition were more than twice as likely to generate hypotheses consistent with the more-accessible schema than the less-accessible schema. If the hypothesis provided to the subjects suggested a schema that was more-accessible, then there was relatively little change in hypothesis generation performance as compared to the "no hypothesis" subjects. If, however, the schema suggested by the hypothesis was less-accessible, and hence less likely to occur to the subjects spontaneously, then there was a dramatic increase in the number of hypotheses generated that were consistent with that schema. There was also a corresponding decrease in hypotheses generated that were consistent with the more-accessible schema. These results are evidence for the biasing effects of schemata.

We also explored a simple technique for reducing the bias. A second group of subjects was given much the same procedure as the first group, except that the subjects who "inherited" hypotheses were asked to generate a hypothesis which was consistent with the data "for another reason." For the subjects who successfully generated such a hypothesis, the bias was practically eliminated. There was an added benefit from this procedure. Less-accessible schemata became more accessible, but the generation of hypotheses consistent with the more-accessible schema was reduced. Possibly subjects had some upper limit to the number of hypothesis that they were willing to generate.

Individual differences in hypothesis generation

We noticed pronounced individual differences in hypothesis-generation ability among our subjects. Some subjects generated more than twice as many hypotheses as a typical subject, and although the typical subject generated impoverished hypothesis sets, there was an occasional exception to this rule.

For practical reasons it might be useful to have a simple means of estimating the hypothesis generation ability of an individual, and the cognitive differences between good and poor hypothesis generators might be enlightening.

Our first study on this topic (5) was fairly traditional. First, we developed criterion measures of hypothesis generation performance. One criterion task was an abstract photo-reconnaissance task where the decision maker was given a simplified copy of a map from the U. S. Census tract. An unknown area was marked on the map, and the subjects' task was to generate as many hypotheses as possible about the identity of this unknown area using the map and several additional items of information. The criterion hypothesis generation score which was finally developed depended on both the quantity and quality of the hypotheses that the subject generated.

Our choice of predictor variables was guided by several considerations. First, the divergent thinking involved in hypothesis generation seemed to be similar to the divergent thinking used in some creative activities. We surveyed this literature and identified several tests that were designed to measure divergent thinking and creativity. These tests were the Alternate Uses test, the Remote Associations test, and a subtest of the AC test of Creative Ability which we called "Possible Reasons." Second, other tests were included to measure such factors as inductive reasoning, and the ability to use the information provided by the tasks.

Alternate Uses was found to be by far the best predictor of hypothesis generation performance (r=.27), but none of the predictors accounted for much of the variance in this ability.

In the second study of this series (5), we took steps to increase the reliability of the criterion measure of hypothesis generation. The Alternate Uses test was retained, and the other tests of creative problem solving were dropped. Tests of general academic achievement (the ACT), and intellectual ability (the Information scale of the WAIS) were added to the battery of predictors. Several different versions of Alternate Uses were also developed to measure possible cognitive skills that might be involved in hypothesis generation.

Our modifications of the Alternate Uses test were based on the following argument. The Alternate Uses test involves generating alternate uses for common household items, such as a coat hanger. Subjects are instructed to generate as many possible uses for a coat hanger as possible. Many of the possible uses for a coat hanger involve using a different schema than "a device for storing clothing in a closet." A coat hanger has many attributes which can be exploited in various ways. It is metal, it conducts electricity, it is ductile, it is long and thin, it is fairly rigid, it doesn't burn at houshold temperatures, etc. The implicit properties of this object could be used as retrieval cues to search memory. Various combinations of these attributes suggest different schemata such as "a device to open a car door" (long, thin, rigid, and ductile), or "marshmallow roaster" (long, thin, rigid and fire resistant). Therefore, a subject who performs well at this task might first analyze an object to determine implicit dimensions or attributes and then use various combinations of these dimensions as retrieval cues for alternate uses. Performance in the Alternate Uses task and in hypothesis generation might have two components, the retrieval of the implicit dimensions and the use of this implicit information to retrieve uses or hypotheses, depending on the task.

With these thoughts in mind, we modified the Alternative Uses test to create two new versions of the test to use in addition to the original version. One of the new versions measured the subjects' ability to retrieve the attributes of the household objects that might be useful retrieval cues, and a second version measured the subjects' ability to generate uses when these attributes or dimensions were explicitly provided by the experimenter.

There were several interesting results from this experiment. First, as has been found in every study dealing with this topic, hypothesis generation of the average subject was impoverished. The mean hypothesis generation score for subjects was about 3 "good" hypotheses per problem, while the lower-bound estimate of the maximum number of logically possible hypotheses was approximately 26 "good" hypotheses and 43 "fair" hypotheses per problem. Second, the correlation between the Alternate Uses test and the criterion measure of hypothesis generation was .51, a considerable gain in predictive power over the previous experiment. This correlation could undoubtedly be increased by item-selection and other methods of test refinement. Such further development could perhaps convert the alternate uses test from a research tool to a useful predictor of hypothesis generation performance. Third, achievement and general intelligence were shown to be only weakly related to hypothesis generation performance.

Both of the proposed cognitive components of hypothesis generation performance were shown to be important. The "retrieval of implicit attributes" component and the "retrieval of hypotheses from attributes" component were significantly related to hypothesis generation performance. An analysis of variance was performed on these data which showed that these two components are additive, uncorrelated factors. Subjects who scored low on both components generated, on the average, 2.15 "good" hypotheses per problem while subjects who scored high on both of these components generated, on the average, 3.6 "good" hypotheses per problem. Of the two components, "retrieval of hypotheses from attributes" accounted for the most variance. This study, therefore, has identified two cognitive skills that appear to be important in hypothesis generation. It also made progress toward the development of a measure of hypothesis-generation ability.

Generalizing to expert populations

Most of our studies employed populations of college students, and the generality of results obtained with this population has been questioned. We deliberately included groups of expert subjects in two studies (4, 7) as a check on the generality of our results obtained with college students. We were interested in determining if experts also generated impoverished hypothesis sets and made excessive plausibility estimates. Our purpose was not to show that expertise has no influence on hypothesis generation. In fact, the hypothesis generation tasks used were carefully chosen so that they could be performed by both college students and expert subjects. Other tasks, requiring the specialized knowlege of an expert, could not be performed by college students, and so were not considered as candidate tasks for these experiments.

Our initial bias was that expert subjects would show considerably different performance than non-experts. Much to our surprise, the experts we studied were quite similar to non-experts in the two performances in which we had the most interest. In the protocol analysis study (7), expert mechanics generated almost exactly the same number of hypotheses as non-experts, and both groups generated impoverished hypothesis sets. The quality of hypothesis sets generated by the experts could not be compared to that of non-experts due to task limitations, but both groups displayed similar excessive plausibility estimates.

Another study (4) was performed which involved expert subjects. This study will be described in more detail below, but the same general conclusions can be reached from this study. The results suggest that observed deficiencies in hypothesis generation can be generalized to experts. We do not claim that expertise is unimportant in hypothesis generation. We do believe, however, that even experts will generate impoverished hypothesis sets and will evaluate these sets as being more exhaustive than they really are. This seems to be the human condition.

Improving Hypothesis Generation

The primary goal of this project was to study the hypothesis generation process, not to find ways of improving hypothesis generation. However, one study was devoted to aiding hypothesis generation. We also discovered several techniques for improving hypothesis generation performance during the course of our research. The study devoted to hypothesis generation aiding and these techniques are described below.

An artificial memory aid for hypothesis generation

Our research suggests that many of the deficiencies in hypothesis generation can be traced to difficulties in retrieving hypotheses from memory. The aiding study (4) employed an artificial memory to aid hypothesis retrieval. Hypotheses retrieved from the artificial memory were displayed to the subjects and they could, if they wished, add these hypotheses to their own set of generated hypotheses. The artificial memory supplemented the hypotheses that subjects were able to retrieve from memory. The aid also exploited the differences between retrieval and recognition in memory. The assumption behind the aid is that subjects may not be able to *retrieve* a plausible hypothesis from memory, but may be able to *recognize* its plausibility if it is presented to them. Thus, the aid was designed to supplement the memory retrieval process.

The artificial memory. Hypothesis generators have used artificial memories of various sorts to aid hypothesis generation. The reference books of a doctor or the maintenance manuals of a mechanic or an electronics technician are examples of artificial memory aids. These aids are primarily useful in situations where routine problems are to be solved. They do not usually suggest hypotheses for rare complexes of symptoms or data. Nevertheless, these artificial memories are so useful that they are often consulted, and when they are unavailable we often deplore their lack. Generally, the information contained in these reference books comes from an authoritative source. This information is so difficult to collect and collate that it usually exists only for commonly-encountered situations.

The problem of constructing an aid to hypothesis retrieval for situations that lack authoritative reference materials is interesting. Consulting an expert would be a possible solution, but we suspect that even experts retrieve incomplete hypothesis sets (7). Several experts might jointly create a more complete hypothesis set if their hypotheses are pooled; this is one reason why doctors often use consultants when making difficult diagnoses. Perhaps one way to achieve a more complete hypothesis set is to pool the hypothesis sets of individuals, as was done in the group research (9).

A difficult problem still remains. The task of creating a pooled hypothesis set for all possible combinations of data or symptoms is impossible for diagnostic situations where many data are present. For example, if there are N data possible, and if a simplifying assumption is made that these data are not mutually exclusive, then the possible number of data complexes is two raised to the Nth power, a potentially large number. Therefore, it may be impossible to convene a panel of experts and to ask them to evaluate every possible data complex that might occur; there may simply be too many possible combinations of data. Perhaps the answer is to use expert judgment to construct an artificial associative memory, and then interrogate this memory to find hypotheses that are suggested by any complex of symptoms or data.

We constructed such an artificial memory. First we asked subjects to generate as many hypotheses as possible for each datum. These hypotheses were pooled across the subjects to create a more-complete hypothesis set than any individual could generate.

These sets were stored in a computer, simulating an associative network. Thus, many plausible hypotheses were associated with each datum in the computer memory. The tagging model developed for modeling human hypothesis retrieval (1) was used to retrieve hypotheses suggested by a complex of data. Hypotheses were tagged in the artificial memory for each datum in the complex, and those hypotheses that received more than a criterion number of tags were retrieved from the artificial memory and displayed to the hypothesis generator.

An evaluation of the artificial memory. A study (4) was performed to evaluate the extent to which this artificial memory aided hypothesis generation. Subjects were given either one or three courses that a student had taken and were asked to generate as many plausible hypotheses as possible about the major of that student. When the subjects finished generating hypotheses, they either started the next problem, or they were shown the results of the search of the artificial memory. This display consisted of a list of hypotheses that had been retrieved from the artificial memory, and the subjects were allowed to add any hypotheses from this list to their hypothesis sets.

There were two groups of subjects. One group consisted of Junior or Senior level students at the University of Oklahoma. The other group was more expert. This group consisted of professional Curriculum Advisors who were employed by the University to give students advice on course offerings and schedule planning.

Performance was measured by calculating the posterior probability of the hypothesis sets that the subjects generated in the aided and unaided conditions. This probability is the probability that the set of generated hypotheses contains the "true" hypothesis. Subjects were told to ignore implausible hypotheses ($P < .02$). For this reason, an optimal hypothesis generator should have generated a hypothesis set that had a probability .889 for the average problem.

The unaided performance of both groups was impoverished. Non-experts had mean hypothesis set probabilities of .477, while experts had mean probabilities of .506. This difference is statistically reliable, but experts performed similarly to non-experts in that both groups generated impoverished hypothesis sets. It will be recalled that a hypothesis set probability was the probability that the true hypothesis was contained in the set of generated hypotheses. An optimal hypothesis generator, one who generated all hypotheses greater than .02, would have a hypothesis set probability of .889.

Both groups increased the plausibility of their hypothesis sets when they used the aid. The non-experts' aided hypothesis sets had a mean probability of .570, while the experts' mean probability was .603. This difference between groups was not reliable, but both groups were aided significantly. The experts showed an improvement of .133, while the non-experts showed an improvement of .185 over their unaided performance. The aid, therefore, provides a noticeable but not dramatic gain in performance.

Perhaps the most interesting result comes from an examination of the hypotheses generated by the subjects and <u>not</u> suggested by the aid. The sum of the posterior probabilities of these hypotheses totaled less than .01. In other words, the aid generated nearly all of the hypotheses that subjects were capable of generating. Had the aid been used as the sole source of generated hypotheses it would have been better than an unaided subject and equivalent to an aided subject. The concept of using an artificial memory to aid hypothesis generation was shown to be viable for those situations where it seems worthwhile to construct such an aid.

Other possiblilities for improving hypothesis generation performance

Some of the results obtained incidentally during our study of the hypothesis generation process might also be usefully employed to improve hypothesis generation. These results will only be mentioned briefly here because they have already been discussed.

Our study of group hypothesis generation strongly suggests that using several hypothesis generators will yield a considerable gain in performance. These results also suggested that social interaction during hypothesis generation reduces performance; the best performance would be achieved by using a synthetic pooling of hypotheses such as that done in the group study (9) and the aiding study (4). Depending upon the importance of the problem, groups of varying sizes can be used, with the pooled hypothesis sets of large groups resulting in a dramatic improvement in performance (9).

If the hypothesis generators are encouraged to try to think of another schema which might explain the data, their hypothesis sets are less biased by hypotheses "inherited" from earlier work on that problem (6). This procedure should be routinely employed as it costs almost nothing to use.

A step which can be taken to reduce the bias in plausibility estimates is to help the hypothesis generator populate the set of unspecified hypotheses (2). Not only does populating the unspecified hypothesis set reduce the bias in plausibility estimates, but it might also be expected to encourage the hypothesis generator to continue to search memory beyond the point where such searches normally stop.

Finally, it seems possible to select good hypothesis generators by means of tests which measure divergent thinking, and our study on this topic (5) suggests that such paper-and-pencil tests might be developed.

Each of these proposed improvements by itself results in a relatively modest gain in performance. If all of these techniques were to be used simultaneously, we would predict that considerable gains in performance might well result.

The "Fat and Happy" Hypothesis Generator

One major conclusion supported by this research is that sets of hypotheses generated by our subjects were impoverished, but subjects estimated that these sets were more complete than they actually were. Similar results have been obtained using a wide variety of tasks, several experimental strategies, and several response modes. Although some variables do effect estimates of the extent of hypothesis generation deficiencies, we have found no exceptions to the general conclusions that subjects generate impoverished hypothesis sets and overestimate their completeness.

During this project we have employed a variety of hypothesis generation tasks, partially to determine if our results were task-specific. We employed tasks where subjects generated hypotheses about the majors of undergraduates, occupations of skilled workmen, and identities of States of the Union (1, 2, 4, 9). Other tasks involved generating the identity of animals (3), and defects in an automobile (7). Two experiments used problems where the object was to generate hypotheses about an unknown geographical area (5, 6). In all of those experiments where a measure of hypothesis generation performance was obtained, subjects generated impoverished hypothesis sets. In all of those experiments where plausibility estimates were obtained, subjects were excessive in their assessments of the completeness of the hypothesis sets.

The same general conclusions that were reached using college students seem to be justified for expert subjects (4, 7). Although this variable was investigated in only two studies, the results suggest that experts and non-experts have similar difficulties.

In one study, it was shown that plausibility estimates were excessive irrespective of whether the subjects were judging hypothesis sets that they had generated or hypothesis sets supplied by the experimenter. In this same study, it was shown that the plausibility estimation measurement technique used in many of these studies produced much the same results as probability estimation.

These results, taken as a whole, present a rather unflattering picture of the hypothesis generator. Hypothesis generators may feel "fat and happy" about the completeness of their hypothesis sets, when the available data about their performance suggests that they should feel "thin and worried." Generated hypothesis sets lack important hypotheses, yet when these sets are evaluated, the hypothesis generator feels that they are more complete that they really are.

Our data suggests that the explanation for the "fat and happy" syndrome lies in deficiences in the memory search process. The subjects' inability to access all plausible hypotheses available in memory seems to be the underlying cause of both poor retrieval from memory and excessive plausibility estimates. The paradox is that these results suggest that hypothesis generators may be unaware of their deficiencies because the difficulty in retrieving hypotheses from memory also affects the evaluative process where they assess the completeness of their performance.

List of Technical Reports with Abstracts

<u>1978</u>

1. Gettys, C., Fisher, S., and Mehle, T. <u>Hypothesis generation and plausibility assessment</u> (Tech. Rep. TR 15-10-78). Norman, Ok.: University of Oklahoma, Decision Processes Laboratory, July 1979.

A hypothesis generation model is described which consists of two subprocesses. Hypotheses are retrieved from memory using several data as retrieval cues in the hypothesis retrieval sub-process. These hypotheses are then evaluated by a plausibility assessment sub-process. Two experiments are described. A memory retrieval experiment examined hypothesis retrieval from memory using multiple data. A memory-tagging model is described which predicts the probability of multi-data hypothesis retrieval. Performance in this task was poor; subjects rarely generated an adequate hypothesis set. A second plausibility assessment experiment was performed where subjects estimated the plausibility of specified hypotheses using varying amounts of data. Plausibility assessments for specified hypotheses were usually extreme in comparison to the posterior odds calculated by Bayes' theorem. This result was also attributed to deficiencies in hypothesis retrieval from memory.

<u>1979</u>

2. Mehle, T., Gettys, C., Manning, C., Baca, S., and Fisher, S. <u>The availability explanation of excessive plausibility assessments</u> (Tech. Rep. TR 30-7-79). Norman, Ok.: University of Oklahoma, Decision Processes Laboratory, July 1979.

The assessment of hypotheses in hypothesis generation involves a comparison between those hypotheses that have been generated (specified) and those that are not generated (unspecified). This study investigated the "availability explanation" (Tversky and Kahneman, 1973) for subjects' overconfidence in estimating the probability of specified hypotheses. The conjecture is that subjects have difficulty retrieving unspecified hypotheses; a complete set of candidate unspecified hypotheses is unavailable during assessment. Therefore, the underpopulated set of unspecified hypotheses is regarded as less probable and the specified set is regarded as more probable. A control group in this study replicated previous findings of overconfidence for specified hypotheses. Two manipulations to increase the availability of unspecified hypotheses were investigated. One manipulation involved explicitly requesting subjects to populate the unspecified set. The other manipulation consisted of computer presentation of candidate unspecified hypotheses. Although in a normative sense, neither manipulation should have affected judgments, results indicated that assessment overconfidence for both experimental groups was reduced. These results support our conjecture that the availability heuristic is at least partially responsible for subjects' excessive behavior in evaluating specified hypotheses.

3. Fisher, S., Gettys, C., Manning, C., Mehle, T., and Baca, S. Consistency checking in hypothesis generation (Tech. Rep. 29-7-79). Norman, Ok.: University of Oklahoma, Decision Processes Laboratory, July 1979.

Three experiments were performed to provide evidence that the generation of hypotheses in response to multiple data may involve two different cognitive processes. First, a candidate hypothesis may be retrieved or activated in memory in response to only part of the available data. This candidate hypothesis may then be checked for consistency against the remaining data. This latter process is called "consistency checking." Experiment 1 was performed to provide evidence that consistency checking occurs during hypothesis generation. Subjects were able to recognize hypotheses which were retrieved during a hypothesis generation problem but not emitted as hypothesis responses, suggesting that consistency checking was responsible for the rejected hypotheses. Experiment 2 indicated that the amount of time needed to process an additional datum in a consistency checking task was less than an estimate of the time needed to process an additional datum in hypothesis retrieval. The results suggest that consistency checking is a high-speed verification process rather than a slower search process. Experiment 3 was performed to provide evidence that consistency checking is a self-terminating process. Subjects' latencies depended upon the position of a disconfirming datum within a data set, supporting this conjecture. The results generally confirmed the existence of a high-speed verification process in hypothesis generation and also suggest that the generation of hypotheses in response to multilple data occurs as a result of dual processes.

4. Gettys, C., Mehle, T., Baca, S., Fisher, S., and Manning, C. A memory retrieval aid for hypothesis generation (Tech. Rep. TR 27-7-79). Norman, Ok.:University of Oklahoma, Decision Processes Laboratory, July 1979.

Hypothesis generation consists of retrieving explanations for data from memory, and assessing these explanations for plausibility. Previous research has established that human hypothesis generation performance is deficient in both hypothesis retrieval and assessment. This study investigates an aid for the hypothesis retrieval process which is based on a model for hypothesis retrieval developed by Gettys, Fisher, and Mehle (1978). A computer simulates the human hypothesis retrieval process by searching an enriched associative memory which contains the associations of a number of individuals in the form of lists of hypotheses for each datum. When the data of a decision problem become known, the appropriate lists are searched by the computer.

Hypotheses that are common to most or all of the lists are suggested to the user, who assesses them for plausibility. An experiment was performed to determine the utility of the aid for both expert and non-expert users. The aid produced a substantial gain in performance for both groups of users, suggesting that further development of the aid would be worthwhile in decision situations which are repeated often enough to warrant the creation of an enhanced artificial memory. Also discussed are several techniques for implementing the aid, and determining the maximum gain in performance that the aid can produce.

5. Manning, C., Gettys, C., Nicewander, A., Fisher, S., and Mehle, T. Predicting individual differences in hypothesis generation (Tech. Rep. TR 28-7-79). Norman, Ok.: University of Oklahoma, Decision Processes Laboratory, July 1979.

Two experiments were performed to determine the extent to which individual differences in hypothesis generation could be predicted. In the first experiment, several published tests of creativity were used as predictors of hypothesis generation ability. The Alternate Uses test was the best predictor of hypothesis generation performance. In a second experiment, measures of achievement, general mental ability, and information were included with Alternate Uses as predictors of performance. Again Alternate Uses was the best predictor of performance. Several variants of the Alternate Uses test were also employed to isolate the components of hypothesis generation. It was found that two components were involved: retrieval of implicit dimensions of the objects and retrieval of uses when the dimensions are explicitly provided. The latter component was found to be by far the most important. It was concluded that good hypothsis generators have skills that enable them to effectively retrieve information stored in memory.

1980

6. Manning, C., and Gettys, C. The effect of a previously-generated hypothesis on hypothesis generation performance (Tech. Rep. TR 8-5-80). Norman, Ok.: University of Oklahoma, Decision Processes Laboratory, August 1980.

An experiment was performed to determine what effects exposure to a previously-generated hypothesis would have on subsequent hypothesis generation. The results showed that hypothesis generation performance is relatively unchanged if the previously-generated hypothesis is consistent with a salient interpretation of the data. However, if the previously-generated hypothesis is consistent with a relatively unusual interpretation of the data, then subjects use both the interpretation that is consistent with the hypothesis and the more commonly used interpretation as cues to retrieve hypotheses. In this case, resulting hypothesis sets included more varied types of hypotheses. Instructions to consider other interpretations of the data also resulted in subjects' generating richer hypothesis sets.

7. Mehle, T. Hypothesis generation in an automobile malfunction inference task (Tech. Rep. TR 25-2-80). Norman, Ok.: University of Oklahoma, Decision Processes Laboratory, February 1980.

Expert and novice subjects generated hypotheses in an automobile troubleshooting inference task. Data collected included subjects' verbal protocols during the inference tasks and subjects' estimates of the probabilities of their generated sets of hypotheses. Analyses indicated that both expert and novice subjects had difficulty generating complete sets of hypotheses and were overconfident in their subjective estimates of the probabilities of generated hypotheses.

8. Casey, J., Mehle, T., and Gettys, C. A partition of group performance into informational and social components in a hypothesis generation task (Tech. Rep. TR 3-3-80) Norman, Ok.: University of Oklahoma, Decision Processes Laboratory, August 1980.

A technique is presented for partitioning group performance into two components: a component due to the increased information possessed by the group and a component representing the change in performance due to social interaction. The hypothesis-generation performance of individuals working alone was compared to the performance of interacting groups of four. The particular task employed permitted calculations of the veridical probabilities of generated sets of hypotheses. Analyses of results were based on a new method, obtained by pooling hypothesis sets from individual subjects to obtain "synthetic" groups. This method permits direct comparisons of interacting and synthetic groups' hypothesis-generation performance. Using this method, we found that groups of four subjects were equivalent to synthetic groups of 1.8 subjects.

9. Gettys, C., Manning, C., Mehle, T., and Fisher, S. Hypothesis generation: A final report of three years of research (TR 15-10-80). Norman, Ok.: University of Oklahoma, Decision Processes Laboratory, October 1980.

This final report summarizes 14 experiments conducted over a three-year period. First discussed is a hypothesis generation model and research which addresses the model. Several major findings were obtained: 1) Hypothesis retrieval from memory is impoverished. Hypothesis generators are not able to retrieve all relevant hypotheses from memory that should be considered in a decision problem. 2) Hypotheses that are retrieved from memory are first checked for logical consistency with the data. Those hypotheses that are logically consistent may be assessed further for plausibility. 3) Hypothesis generators think that collections of hypotheses which they generated are much more complete than they actually are.

The next section discusses research on hypothesis generation performance. Topics include protocol analysis, group hypothesis generation, the biasing effects of schemata, individual differences in hypothesis generation, and generalizing to expert populations.

A third section is devoted to a survey of research relevant to aiding the hypothesis generation process. An artificial aid for retrieving hypotheses from memory is discussed. Also discussed are other ways of improving hypothesis generation performance.

The general conclusion of this project is that both the failure to retrieve enough hypotheses from memory and the subjects' belief that these collections of hypotheses are more complete than they actually are can be traced to deficiencies in the memory retrieval process.

OFFICE OF NAVAL RESEARCH, CODE 455
TECHNICAL REPORTS DISTRIBUTION LIST

CDR Paul R. Chatelier
Office of the Deputy Under Secretary
of Defense
OUSDRE (E&LS)
Pentagon, Room 3D129
Washington, D.C. 20301

Director
Engineering Psychology Programs
Code 455
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217 (5 cys)

Director
Analysis and Support Division
Code 230
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Director
Naval Analysis Programs
Code 431
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Director
Operations Research Programs
Code 434
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Director Statistics and Probability
Program
Code 436
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Director
Information Systems Program
Code 437
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Special Assistant for Marine
Corps Matters
Code 100M
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Commanding Officer
ONR Branch Office
ATTN: Dr. J. Lester
Building 114, Section D
666 Summer Street
Boston, Massachusetts 02210

Commanding Officer
ONR Branch Office
ATTN: Dr. C. Davis
536 South Clark Street
Chicago, Illinois 60605

Commanding Officer
ONR Branch Office
ATTN: Mr. R. Lawson
1030 East Green Street
Pasadena, California 91106

Office of Naval Research
Scientific Liaison Group
American Embassy, Room A-407
APO San Francisco, California 96503

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, D.C. 20375 (6 cys)

Dr. Robert G. Smith
Office of the Chief of Naval
Operations, OP987H
Personnel Logistics Plans
Washington, D.C. 20350

Naval Training Equipment Center
ATTN: Technical Library
Orlando, Florida 32813

ONR, Code 455, Technical Reports Distribution List

Human Factors Department
Code N215
Naval Training Equipment Center
Orlando, Florida 32813

Dr. Alfred F. Smode
Training Analysis and Evaluation Group
Naval Training Equipment Center
Code N-00T
Orlando, Florida 32813

Scientific Advisor to DCNO (MPT)
OP 01T (Dr. Marshall)
Washington, D.C. 20370

CDR Thomas Berghage
Naval Health Research Center
San Diego, California 92152

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Groton, Connecticut 06340

Navy Personnel Research and Development Center
Manned Systems Design, Code 311
San Diego, California 92152

Navy Personnel Research and Development Center
Code 305
San Diego, California 92152

Navy Personnel Research and Development Center
Management Support Department
Code 210
San Diego, California 92151

CDR P. M. Curran
Code 604
Human Factors Engineering Division
Naval Air Development Center
Warminster, Pennsylvania 18974

Dean of the Academic Departments
U.S. Naval Academy
Annapolis, Maryland 21402

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, California 93940

Dean of Research Administration
Naval Postgraduate School
Monterey, California 93940

Mr. Warren Lewis
Human Engineering Branch
Code 8231
Naval Ocean Systems Center
San Diego, California 92152

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, D.C. 20380

Mr. Arnold Rubinstein
Naval Material Command
NAVMAT 08D22
Washington, D.C. 20360

Mr. Phillip Andrews
Naval Sea Systems Command
NAVSEA 0341
Washington, D.C. 20362

Naval Sea Systems Command
Personnel & Training Analyses Office
NAVSEA 074C1
Washington, D.C. 20362

LCDR W. Moroney
Code 55MP
Naval Postgraduate School
Monterey, California 93940

Mr. Merlin Malehorn
Office of the Chief of Naval Operations (OP 102)
Washington, D.C. 20350

Mr. J. Barber
HQS, Department of the Army
DAPE-MBR
Washington, D.C. 20310

ONR, Code 455, Technical Reports Distribution List

Dr. Joseph Zeidner
Technical Director
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, Virginia 22333

Director, Organizations and
Systems Research Laboratory
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, Virginia 22333

Dr. Edgar M. Johnson
Organizations and Systems Research
Laboratory
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, Virginia 22333

Technical Director
U.S. Army Human Engineering Labs
Aberdeen Proving Ground, MD 21005

U.S. Army Aeromedical Research Lab
ATTN: CPT Gerald P. Krueger
Ft. Rucker, Alabama 36362

ARI Field Unit-USAREUR
ATTN: Library
C/O ODCSPER
HQ USAREUR & 7th Army
APO New York 09403

U.S. Air Force Office of Scientific
Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, D.C. 20332

Dr. Donald A. Topmiller
Chief, Systems Engineering Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, Ohio 45433

Air University Library
Maxwell Air Force Base, Alabama 36112

Dr. Gordon Eckstrand
AFHRL/ASM
Wright-Patterson AFB, Ohio 45433

North East London Polytechnic
The Charles Myers Library
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

Professor Dr. Carl Graf Hoyos
Institute for Psychology
Technical University
8000 Munich
Arcisstr 21
FEDERAL REPUBLIC OF GERMANY

Dr. Kenneth Gardner
Applied Psychology Unit
Admiralty Marine Technology
Establishment
Teddington, Middlesex TW11 OLN
ENGLAND

Director, Human Factors Wing
Defense & Civil Institute of
Environmental Medicine
Post Office Box 2000
Downsview, Ontario M3M 3B9
CANADA

Dr. A. D. Baddeley
Director, Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF
ENGLAND

Defense Documentation Center
Cameron Station, Bldg. 5
Alexandria, Virginia 22314 (12 cys)

Dr. Craig Fields
Director, Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, Virginia 22209

Dr. Judith Daly
Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, Virginia 22209

ONR, Code 455, Technical Reports Distribution List

Dr. Stanley Deutsch
Office of Life Sciences
National Aeronautics and Space Administration
600 Independence Avenue
Washington, D.C. 20546

Professor Douglas E. Hunter
Defense Intelligence School
Washington, D.C. 20374

Dr. Robert R. Mackie
Human Factors Research, Inc.
5775 Dawson Avenue
Goleta, California 93017

Dr. Gary McClelland
Institute of Behavioral Sciences
University of Colorado
Boulder, Colorado 80309

Human Resources Research Office
300 N. Washington Street
Alexandria, Virginia 22314

Dr. Miley Merkhofer
Stanford Research Institute
Decision Analysis Group
Melo Park, California 94025

Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army-Navy Drive
Arlington, Virginia 22202

Professor Judea Pearl
Engineering Systems Department
University of California-Los Angeles
405 Hilgard Avenue
Los Angeles, California 90024

Professor Howard Raiffa
Graduate School of Business Administration
Harvard University
Soldiers Field Road
Boston, Massachusetts 02163

Dr. Arthur I. Siegel
Applied Psychological Services, Inc.
404 East Lancaster Street
Wayne, Pennsylvania 19087

Dr. Paul Slovic
Decision Research
1201 Oak Street
Eugene, Oregon 97401

Dr. Amos Tversky
Department of Psychology
Stanford University
Stanford, California 94305

Dr. Gershon Weltman
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, California 91364

Dr. Meredith P. Crawford
American Psychological Association
Office of Educational Affairs
1200 17th Street, NW
Washington, D.C. 20036

Dr. Ward Edwards
Director, Social Science Research Institute
Unversity of Southern California
Los Angeles, California 90007

Dr. Kenneth Hammond
Institute of Behavioral Science
University of Colorado
Room 201
Boulder, Colorado 80309

Dr. William Howell
Department of Psychology
Rice University
Houston, Texas 77001

Journal Supplement Abstract Service
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036 (3 cys)

Dr. Clinton Kelly
Decisions and Designs, Inc.
8400 Westpark Drive, Suite 600
P.O. Box 907
McLean, Virginia 22101

Mr. Richard J. Heuer, Jr.
27585 Via Sereno
Carmel, California 93923

ONR, Code 455, Technical Reports Distribution List

Mr. Tim Gilbert
The MITRE Corporation
1820 Dolly Madison Blvd.
McLean, Virginia 22102